Cluster Analysis of Stocks

Team Members: Jordan Zane, Caleb Clough, Matt Dyer, Evan Linden

Description

This project will aim to mine extensive longitudinal American stock market data to identify interesting trends. In particular, our group aims to cluster individual stocks into groups to see what individual stocks tend to “move” together. Ideally this analysis will uncover greater nuance and cross relationships outside of typical stock groupings (technology, manufacturing, finance, etc.) and allow for novel (or atleast modified) classification grouping schemes.

Prior Work

From Kaggle:

- Which stocks should I invest in? - included daily percentage change in price and volume, Correlation of percent change, Predicting expected gain and loss for each ticker, and growth trends of tickers
- Stock diversity analysis - analysing clusters to identify stocks with opposite trends
- Linear regression of stock histories - using a regression model to predict stock prices
- Wsb stock analysis - using popular tickers mentioned on reddit.com to analyze growth

Datasets

Our main dataset is a .csv containing 5 years of AMEX, NYSE, and NASDAQ end of day data. The source of this data is Kaggle and the URL is as follows:

https://www.kaggle.com/qks1lver/amex-nyse-nasdaq-stock-histories?select=fh_5yrs.csv

Additionally, if this first dataset appears inadequate, we have identified a second dataset of end-of-day stock information from Kaggle that could possibly be integrated into a final, larger dataset if necessary. URL is as follows:

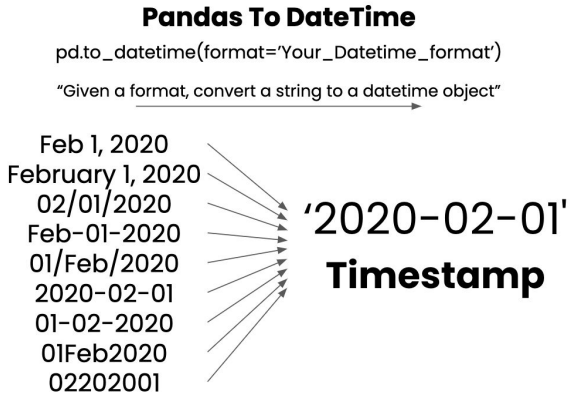
https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs

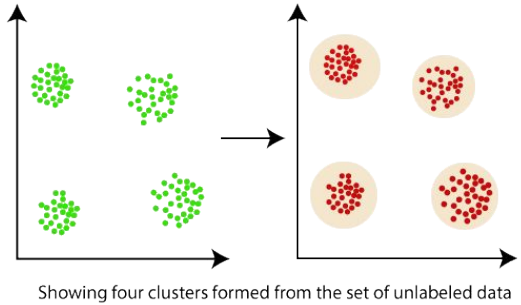
Both sets of data have been frozen, downloaded, and stored on the cloud at the following location:

https://drive.google.com/drive/folders/17RyZ9pbuVMTZXn2SvbAhO8fvzA-08Kbv?usp=sharing

Proposed Work

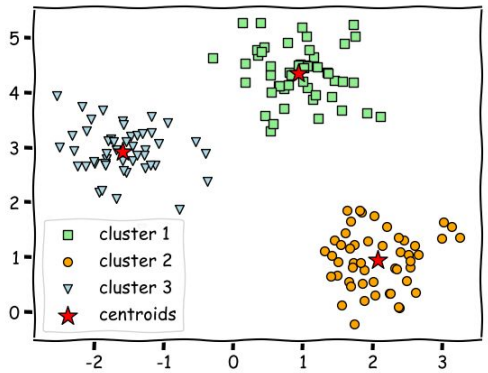
- Data Cleaning
  - Drop any NULL data values, if there are any
  - Replace any string values from the data set (e.g. $)
  - Convert Dates using Pandas to_datetime
  - Examine for any outliers and remove if necessary
- Data Preprocessing
  - Normalize data using min/max normalization
  - Reduce the size of the dataset using a data reduction technique
    - Use normalized data to implement clustering of Stocks
    - Implement data cube

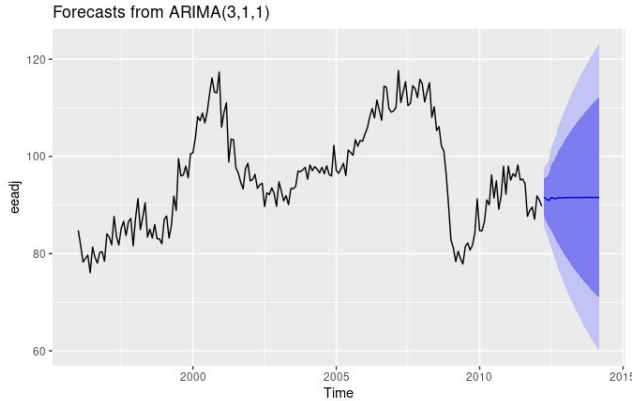




Proposed Work Cont'd

- Data Integration
  - Kaggle dataset updated to 6/14/20
  - All subsequent dates to current date scraped from Yahoo Finance
  - Use stock symbols (eg. TSLA) for entity identification
- Data Analysis
  - Use daily stock movement (close - open)
  - Clustering to identify similar stocks
    - K-means clustering
    - Identify possible outliers, use K-medoids if necessary
  - Time series analysis to forecast cluster trends
    - ARIMA models
    - Exponential smoothing





List of Tools

- Jupyter Notebook
- Python
- NumPy
- Matplotlib
- Pandas
- Seaborn
- Sklearn
- GitHub
  - https://github.com/evanlinden/Data-Mining-Group-02.git

Evaluation

- Stock metadata intuition (eg Are tech companies is same clusters?)
- Split data into testing set (last 1/2 years) and training set (all other years)
- Assess similarity in stock movement among clusters and between clusters
  - Move in same direction?
  - Move at same magnitude?
- Inertia score in k-means clustering
- Elbow method for optimal number of clusters